Sparse Penalty in Deep Belief Networks: Using the Mixed Norm Constraint

Xanadu C. Halkias
DYNI, LSIS, Universitè du Sud,
Avenue de l'Université - BP20132, 83957 LA GARDE CEDEX - FRANCE
xanadu.halkias@univ-tln.fr

Sébastien Paris
DYNI, LSIS CNRS UMR 7296, Aix-Marseille University Domaine universitaire de Saint Jérôme
Avenue Escadrille Normandie Niemen,13397 MARSEILLE Cedex 20, FRANCE
sebastien.paris@lsis.org

Hervé Glotin
DYNI, LSIS CNRS UMR 7296, Université Sud Toulon-Var, Institut Universitaire de France
Avenue de l'Université - BP20132, 83957 LA GARDE CEDEX - FRANCE
glotin@univ-tln.fr

Abstract

Deep Belief Networks (DBN) have been successfully applied on popular machine learning tasks. Specifically, when applied on hand-written digit recognition, DBNs have achieved approximate accuracy rates of 98.8%. In an effort to optimize the data representation achieved by the DBN and maximize their descriptive power, recent advances have focused on inducing sparse constraints at each layer of the DBN. In this paper we present a theoretical approach for sparse constraints in the DBN using the mixed norm for both non-overlapping and overlapping groups. We explore how these constraints affect the classification accuracy for digit recognition in three different datasets (MNIST, USPS, RIMES) and provide initial estimations of their usefulness by altering different parameters such as the group size and overlap percentage.

1 Introduction

Restricted Boltzmann Machines (RBMs) are Energy Based Models (EBMs) that have been extensively used for a diverse set of machine learning applications mainly due to their generative and unsupervised learning framework. These applications range from image scene recognition and generation [**?**], video-sequence recognition [**?**] and dimensionality reduction [**?**].

An equally important aspect of RBMs is that they serve as the building blocks of DBNs [**?**]. Their use as such has been favored in the machine learning community due to the conditional independence between the hidden units in the RBM that allows for the efficient and computationally tractable implementation of deep architectures.

In recent years, sparsity has become an important requirement in both shallow [add cite] and deep architectures. Although primarily used in statistics for optimization tasks in order to overcome the curse of dimensionality in various applications, it also serves as a way to emulate biologically plausible models of the human visual cortex, where it has been shown that sparsity is an integral process in the hierarchical processing of visual information [**? ? ?**].

Moreover, an added benefit of using sparse constraints in the form of mixed norm regularizers in deep architectures is that they can alleviate their restrictive nature by allowing implicit interactions between the hidden units in the RBMs. Mixed norm regularizers such as $l_{1,2}$ have been extensively used in statistics and machine learning [**?**]. In this paper we provide initial results when inducing sparse constraints by using a mixed norm regularizer on the activation probabilities of the RBMs. The mixed norm is applied on both non-overlapping and overlapping groups. We also show that this regularizer can be used to train DBNs, and offer results for the task of digit recognition using several datasets.

2 Restricted Boltzmann Machines

An RBM is a type of two layer neural network comprised of a visible layer that represents the observed data x and a hidden layer that represents the hidden variables h. The addition of these hidden units allows the model an increased capacity in expressing the underlying distribution of the observed data.

RBMs are energy based models and as such they define a probability distribution through an energy function as seen in Eq. 1

$$p(\mathbf{x}, \mathbf{h}) = \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z} \tag{1}$$

Where Z, provided in Eq. 2, is called the partition function and is a normalizing factor ensuring that Eq. 1 is a probability.

$$Z = \sum_{\mathbf{x},\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})} \tag{2}$$

In the case of an RBM the energy function $E(\mathbf{x}, \mathbf{h})$ is defined in Eq. 3:

$$E_\theta(\mathbf{x}, \mathbf{h}) = -\sum_{i=1}^{I}\sum_{j=1}^{J} x_i h_j w_{ij} - \sum_{i=1}^{I} b_i x_i - \sum_{j=1}^{J} a_j h_j, \tag{3}$$

$\mathbf{b}$ are the visible unit biases and $\boldsymbol{a}$ are the hidden unit biases.

In the common case where we are using stochastic binary units for both visible and hidden units, then the conditional probabilities of activation are obtained by:

$$\begin{aligned} p(x_i = 1|\mathbf{h}) &= \sigma(b_i + \textstyle\sum_j h_j w_{ij}) \\ p(h_j = 1|\mathbf{x}) &= \sigma(a_j + \textstyle\sum_i x_i w_{ij}), \end{aligned} \tag{4}$$

where σ is the sigmoid function and

$$\sigma(f(x)) \triangleq \frac{1}{1 + e^{-f(x)}}. \tag{5}$$

Since an RBM does not allow for connections amongst hidden units or amongst visible units we can easily obtain Eq. 6.

$$\begin{aligned} p(\mathbf{x}|\mathbf{h}) &= \prod_i p(x_i|\mathbf{h}) \\ p(\mathbf{h}|\mathbf{x}) &= \prod_j p(h_j|\mathbf{x}) \end{aligned} \tag{6}$$

Intuitively, the observed data, $\mathbf{x}$ will be modeled by those hidden units, $\mathbf{h}$ that are expressed with a high conditional probability $p(h_j|\mathbf{x})$. The goal of adding sparse constraints to the network is to allow for the salient activation of the hidden units based on the differences of the observed data. As a result, we can achieve an initial clustering of the observed data that will increase the discriminative power of the model.

2.1 Training an RBM

RBMs are energy based, generative models that are trained to model the marginal probability $p(\boldsymbol{x})$ of the observed data where:

$$p(\mathbf{x}) = \sum_{\mathbf{h}\in\{0,1\}^J} p(\mathbf{x}, \mathbf{h}). \tag{7}$$

In general, energy based models can be learnt by performing gradient descent on the negative log-likelihood of the observed data. Specifically, to learn the parameters of the network we need to compute the gradient provided in Eq. 8 given the observed (training) data $\boldsymbol{x^l}$.

$$-\frac{\partial logp(\mathbf{x})}{\partial \boldsymbol{\theta}} = \langle \frac{\partial E_\theta(\boldsymbol{x^l}, \mathbf{h})}{\partial \boldsymbol{\theta}} \rangle_{\boldsymbol{h}} - \langle \frac{\partial E_\theta(\mathbf{x}, \mathbf{h})}{\partial \boldsymbol{\theta}} \rangle_{\boldsymbol{x},\boldsymbol{h}}, \tag{8}$$

where $\langle \cdot \rangle_n$ denotes the expectation with respect to n. As evident in Eq.8, the gradient has two phases. The positive phase which tries to lower the energy of the training data $\boldsymbol{x^l}$ and the negative phase which tries to increase the energy of all x in the model.
Assessing the energy on all the data can be an intractable task given the size of the network and the number of possible configurations. In order to obtain an approximation Hinton (2006) successfully proposed the use of Contrastive Divergence (CD). This allows us to sample an approximation of the expectation over $(\mathbf{x}, \mathbf{h})$ using Gibbs sampling at only k steps. Empirically, it has been shown that setting $k = 1$ will provide an adequate approximation although it will not follow the theoretical gradient [**?**].

Applying CD on Eq. 8 we can obtain the following update equations for the parameters of the network.

$$\Delta \boldsymbol{w}_{\cdot j} = \tfrac{1}{L} \sum_{l=1}^{L} \boldsymbol{x^l} p(h_j = 1|\boldsymbol{x^l}) - \widetilde{\boldsymbol{x}}^{\boldsymbol{l}} p(h_j = 1|\widetilde{\boldsymbol{x}}^{\boldsymbol{l}}) \tag{9}$$

$$\Delta b_i = \tfrac{1}{L} \sum_{l=1}^{L} p(x_i^l = 1|\boldsymbol{h}) - p(\widetilde{x}_i^l = 1|\boldsymbol{h}) \tag{10}$$

$$\Delta a_j = \tfrac{1}{L} \sum_{l=1}^{L} p(h_j = 1|\boldsymbol{x^l}) - p(h_j = 1|\widetilde{\boldsymbol{x}}^{\boldsymbol{l}}), \tag{11}$$

where the $\widetilde{(\cdot)}$ defines the generated distributions obtained by the CD.

In the next section, we introduce a general version of sparse constraints in the learning phase of the RBM through the use of the mixed norm in an effort to control the activation probabilities of the hidden units.

3 Mixed Norm RBMs

Several attempts in inducing sparse constraints in the RBM by [**?** **?**] have been successful in increasing the discriminative power of the models. Examples of these sparse constraints range from weight decay [**?**] to modified norm penalties [**?**]. In this paper we focus on the generalized penalty of the mixed norm ($l_{1,2}$), but will also provide a theoretical and practical implementation for the use of overlapping groups. We will refer to this generalized penalty applied to the expectations of the activation probabilities as the Mixed Norm RBM (MNRBM).

As mentioned before, learning an RBM consists of performing gradient descent on the negative log-likelihood. We can thus define the cost function L to be minimized as $L = -logp(\boldsymbol{x})$. When applying the mixed norm regularizer the cost function takes the general form of Eq. 12.

$$L = -logp(\mathbf{x}) + \lambda(\|p(\mathbf{h} = 1|\mathbf{x})\|_{1,2}) \tag{12}$$

Where λ is a regularizer constant. The second term of Eq. 12 defines the mixed norm penalty on the expectations of the hidden unit activation probabilities. In order to apply the mixed norm we assume that the hidden units are divided into groups. These groups can be non-overlapping or overlapping. As a result, we are able to penalize a whole group and not just individual hidden units.

MNRBM with non-overlapping groups: Given an RBM with J hidden units we define a partition of the hidden units into groups P_m where $m = 1, 2, ...M$. The groups are non-overlapping and of equal size to alleviate computational issues. The mixed norm penalty for a data sample $\boldsymbol{x^l}$ is defined in Eq. 14.

$$\begin{aligned} \|p(\mathbf{h} = 1|\boldsymbol{x^l})\|_{1,2} &= \sum_{m=1}^{M} \|p(\boldsymbol{P_m}|\boldsymbol{x^l})\|_2 \\ &= \sum_{m=1}^{M} \sqrt{\sum_{k \in P_m} p(h_k = 1|\boldsymbol{x^l})^2} \end{aligned} \tag{13}$$

In practice, the desire behind the application of the mixed norm penalty is to set groups of the hidden units to zero when representing the observed data by forcing their activation probabilities to zero. As a result, given an observed data sample only a small number of groups of hidden units will be activated, leading to its sparse representation.

MNRBM with overlapping groups: Given an RBM with J hidden units we define a partition of the hidden units into groups P_m where $m = 1, 2, ...M$. The groups are overlapping and of equal size. Depending on the percentage of overlap, a we will obtain a new set of groups P'_k where $k = 1, 2, ...K$.
We can then define a set of augmented hidden units $J' = \{h' \in J : \forall P_k, P_k \bigcup P_m = P_m, J' \supset J\}$. Subsequently, given an RBM with J' hidden units, we can then consider that the set P'_k, defines non-overlapping, equally sized groups [**?**]. The mixed norm penalty for a data sample $\boldsymbol{x^l}$ is defined in a similar way as in Eq. 14.

$$\begin{aligned} \|p(\boldsymbol{h'} = 1|\boldsymbol{x^l})\|_{1,2} &= \sum_{k=1}^{K} \|p(\boldsymbol{h'}_m|\boldsymbol{x^l})\|_2 \\ &= \sum_{k=1}^{K} \sqrt{\sum_{k \in P'_k} p(h'_k = 1|\boldsymbol{x^l})^2} \end{aligned} \tag{14}$$

3.1 Training the Mixed Norm RBM

In order to train the MNRBM with non-overlapping groups and obtain the model parameters $\boldsymbol{\theta}$ we need to minimize the cost function presented in Eq. 12. This can be achieved by performing a coordinate descent once we have obtained the gradients of the regularizers.

The gradient of the mixed norm penalty for the weights, W is as follows:

$$\begin{aligned} &\tfrac{\partial}{\partial \boldsymbol{w}_{\cdot j}}(\|p(\mathbf{h} = 1|\boldsymbol{x^l})\|_{1,2}) = \\ &= \tfrac{1}{2} \cdot \frac{1}{\sqrt{\sum\limits_{k \in P_m} p(h_k=1|\boldsymbol{x^l})^2}} \cdot 2 \cdot p(h_k = 1|\boldsymbol{x^l}) \cdot \tfrac{\partial p(h_k=1|\boldsymbol{x^l})}{\partial \boldsymbol{w}_{\cdot j}} \\ &= \tfrac{p(h_k=1|\boldsymbol{x^l})}{\|p(\boldsymbol{h}_m|\boldsymbol{x^l})\|_2} \cdot \tfrac{\partial p(\boldsymbol{h}_k=1|\boldsymbol{x^l})}{\partial w_{\cdot j}} \\ &= \tfrac{p(h_k=1|\boldsymbol{x^l})}{\|p(\boldsymbol{h}_m|\boldsymbol{x^l})\|_2} \cdot p(h_k = 1|\boldsymbol{x^l})[1 - p(h_k = 1|\boldsymbol{x^l})] \cdot \boldsymbol{x^l} \\ &= \tfrac{p(h_k=1|\boldsymbol{x^l})^2}{\|p(\boldsymbol{h}_m|\boldsymbol{x^l})\|_2} \cdot p(h_k = 0|\boldsymbol{x^l}) \cdot \boldsymbol{x^l}. \end{aligned} \tag{15}$$

When applied on the expectations of the activation probabilities the mixed norm penalty will follow their trend while forcing the groups that include members with low activation probabilities towards zero. The l_2 norm in the denominator ensures that the groups with low activations will be pushed further closer to zero.

Given the gradients of the penalties the update equations for the MNRBM are presented bellow:

$$\begin{aligned} &\Delta \boldsymbol{w}_{\cdot j} = \\ &\tfrac{1}{L} \sum_{l=1}^{L} [(p(h_j = 1|\boldsymbol{x^l}) + \lambda \tfrac{p(h_j=1|\boldsymbol{x^l})p(h_j=0|\boldsymbol{x^l})}{\sqrt{\sum p(h_m=1|\boldsymbol{x^l})^2}}) \cdot \boldsymbol{x^l} - p(h_j = 1|\widetilde{\boldsymbol{x}}^{\boldsymbol{l}})\widetilde{\boldsymbol{x}}^{\boldsymbol{l}}] \end{aligned} \tag{16}$$

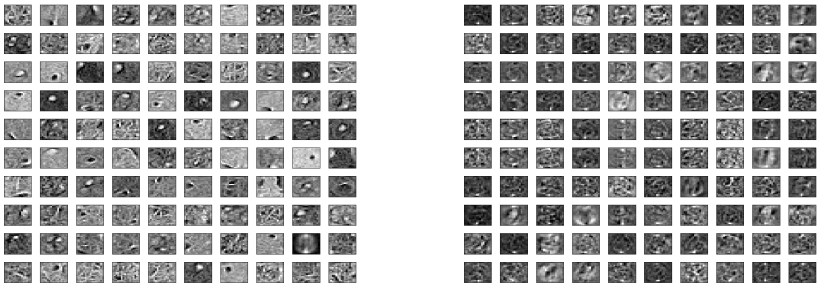
$$\begin{aligned} &\Delta a_j = \\ &\tfrac{1}{L} \sum_{l=1}^{L} [(p(h_j = 1|\boldsymbol{x^l}) + \lambda \tfrac{p(h_j=1|\boldsymbol{x^l})p(h_j=0|\boldsymbol{x^l})}{\sqrt{\sum p(h_m=1|\boldsymbol{x^l})^2}} - p(h_j = 1|\widetilde{\boldsymbol{x}}^{\boldsymbol{l}})] \end{aligned} \tag{17}$$

The detailed steps for training the MNRBM are depicted in Algorithm 1.

The general penalty of Eq. 12 allows us through the manipulation of the constant regularizer, λ, the group size and percentage of overlap to obtain different types of architectures. In this case, the sparsity is induced at the group level of the hidden units whereby the observed data is represented by a small number of groups of hidden units. The λ constant is empirically determined [**?**] based on the task at hand.

Algorithm 1 Mixed Norm RBM learning algorithm

1. Update the parameters $\boldsymbol{\theta}$ using CD and Eq. 9- 11
2. Update the parameters again using the gradient of the regularizations as in Eq. 16- 17
3. Repeat steps 1, 2 until convergence



(a) Mixed norm RBM non-overlapping groups (group size = 20)

(b) Mixed norm RBM overlapping groups (group size = 50, overlap = 20%)

Figure 1: Sample learned weights W for the mixed norm RBM using the USPS data set

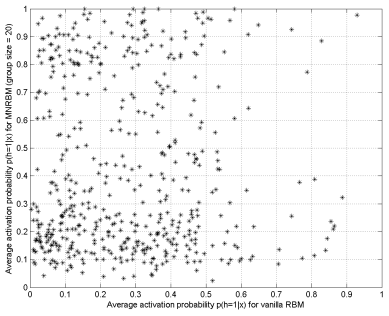
Fig. 1 shows sample weights for the mixed norm RBM when using $\lambda = 0.1$ for both non-overlapping and overlapping groups. Fig 2 provides the average probability activations for the hidden units given a batch of the USPS training data. As seen in the figure, the activation probabilities of the hidden units appear to be more towards the left-hand side of the figure which is the desired effect. However, there appears to be a bimodality whereby a large proportion of the activation probabilities is set to a high value for the non-overlapping groups MNRBM. This may be attributed to the choice and size of groups when applying the mixed norm penalty. Given that the activation probabilities are pushed towards high values one can expect that such a process may have an adverse result for classification tasks since the hidden units will over-represent the observed data. However, in the case of the overlapping groups most of the activations are pushed towards zero. Although, that is the goal of adding sparse constraints is to force the activations to zero, in this case we may be dealing with a biased system that actually under represents the data distribution.

3.2 Data

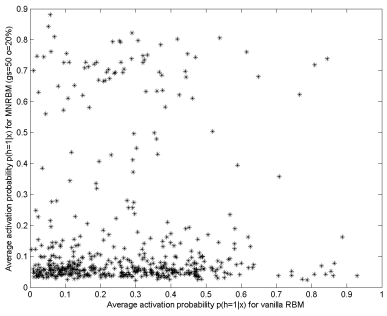
We have used three different data sets in order to train and test the network.

- MNIST is a popular data set in the community for hand-written digit recognition and is comprised of 70000, 28×28 images (60000 train - 10000 test). It is publicly available at *yann.lecun.com/exdb/mnist*.
- The RIMES data set which was created by asking volunteers to write hand written letters for different scenarios. In this paper we used the digit set of the data base. In total the set we used was comprised of 37200 images of different sizes (29800 train - 7400 test). Further information can be obtained at *www.rimes-database.fr*.
- The USPS digit data set that we used is comprised of 9280 (7280 train - 2000 test), 16×16 images. The extracted images were scanned from mail in working U.S. Post Offices [**?**].

In order to achieve the cross-training and testing all images were resized to have the same size as the MNIST dataset (28×28) given its extensive use in this task. All images were also checked to ensure that orientations/translations were uniform across the data sets. No other pre-processing was employed. Example images from the three datasets can be seen in figure 3



(a) Mixed norm RBM non-overlapping groups (group size = 20)



(b) Mixed norm RBM overlapping groups (group size = 50, overlap = 20%)

Figure 2: Average of hidden unit activation probabilities for the mixed norm RBM using a batch of the USPS data set. Y-axis: Hidden unit activation probabilities for mixed norm RBM. X-axis: Hidden unit activations for vanilla RBM

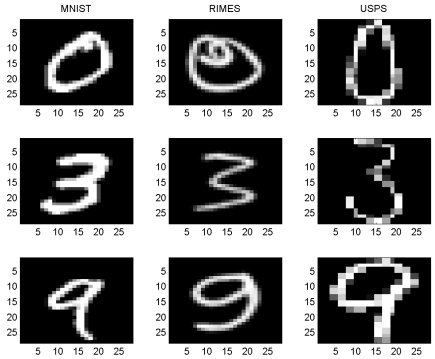


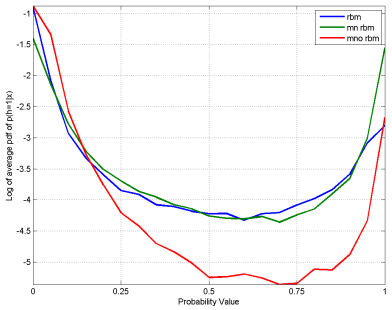
Figure 3: Examples of images from the three datasets MNIST (left), RIMES (center) and USPS (right)

3.3 Experimental Results: Pre-training DBNs with MNBMs

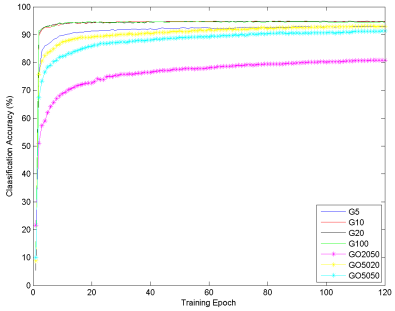
RBMs became increasingly popular when Hinton and Salakhudinov [**?**] [**?**] used them as building blocks for creating and pre-training efficient DBNs. The proposed MNRBMs can be utilized in the same manner to initialize DBNs and obtain a sparse and computationally efficient representation of the observed data.

In order to offer a comparative view between the different architectures we used Hinton's model for digit recognition, but we substituted the vanilla RBM with the proposed MNRBM. We pre-trained a $500 - 500 - 2000$ DBN and tested it on three different data sets, MNIST, RIMES and USPS.

Continuing, to obtain classification error rates we added 10 softmax layers to get the posterior probabilities for the different classes. The network was fine-tuned using conjugate gradient as described in [**?**]. The constant regularizer was empirically set to $\lambda = 0.1$ for all different models [**?**]. Continuing, for the mixed norm architecture with non-overlapping groups we used different group sizes for the hidden units, 5, 10, 20 and 100 respectively. In the case of overlapping groups we used group sizes of 20, and 50 with $a = \{20\%, 50\%\}$. Results on the classification accuracy and the computational cost of the models can be seen in Table 1 and Table 2 respectively. All experiments were performed on a 24 core server (AMD Opteron processor 8435) with a core CPU of 2.6GHz and a cache of 512KB.



(a) Average Probability density function for activation probabilities, $p(\boldsymbol{h} = \mathbf{1}|\boldsymbol{x})$



(b) Classification Accuracy for the different penalties using the USPS data set

Figure 4: Average log pdf for activation probabilities for the vanilla RBM and mixed norm RBM using a batch of the USPS data set and classification accuracy for the USPS data set using the different architectures

Table 1: Classification accuracies for the different architectures based on the general sparse penalty.

ARCHITECTURE	MNIST	RIMES	UPS
DBN	**98.83%**	99.30%	**94.85%**
MN DBN (5)	97.28%	99.24%	92.90%
MN DBN (10)	**98.83%**	99.33%	94.70%
MN DBN (20)	98.77%	99.38%	94.65%
MN DBN (100)	98.80%	**99.40%**	94.35%
MN W/O DBN (20/20%)	95.10%	95.70%	85.05%
MN W/O DBN (20/50%)	93.50%	93.62%	80.90%
MN W/O DBN (50/20%)	96.50%	97.60%	92.95%
MN W/O DBN (50/50%)	95.84%	96.27%	91.35%

From Table 1 we can infer that our proposed mixed norm penalty can offer the flexibility of creating architectures that will be able to matchthe classification accuracy of the models depending on the underlying distributions. It appears that for the task of hand-written digit recognition the distribution of the observed data favors the use of larger non-overlapping group sizes for the mixed norm architectures.

In order to get a better understanding of the impact of the different sparse constraints and architectures, Figure 4 depicts the average probability density functions of the expectations of the activation probabilities for the MNIST training data.

It is interesting to note that the proposed architectures that utilize the mixed norm penalty (MNDBN)with the overlapping groups tend to aggressively push their activation probabilities to zero. However, these architectures also tend to offer lower accuracy rates which can be attributed to an inability of the models to concisely capture the underlying data. A possible way for exploring this phenomenon further may be to constrain the penalty of the expectations as seen in [**?**].

3.4 Conclusions

In this work we provided some first insights for the use of the mixed norm sparse constraint in DBNs. We performed experiments using three different data sets for the task of hand written digit recognition and offered a practical approach for the use of overlapping groups with the mixed norm constraint. Although, our initial experiments were limited in the use of equal size overlapping groups, one could easily extended to non-symmetric overlapping groups, using a similar methodology. Inducing sparse constraints based on specific geometries may also provide better results in the case of digit recognition and offer more interesting results for tasks such as scene categorization.

Table 2: CPU times for the different architectures based on the general sparse penalty.

ARCHITECTURE	MNIST	RIMES	UPS
DBN	167.90H	>60H	31.15H
MN DBN (5)	62.14H	33.70H	8.62H
MN DBN (10)	66.10H	40.70H	10.00H
MN DBN (20)	70.10H	69.80H	12.75H
MN DBN (100)	71.50H	85.80H	15.85H
MN W/O DBN (20/20%)	>60H	39.27H	10.40H
MN W/O DBN (20/50%)	>60H	>45H	22.90H
MN W/O DBN (50/20%)	>60H	35.60H	9.56H
MN W/O DBN (50/50%)	>70H	>45H	24.00H